Group Project – Flint Water Line Replacements 2016

DS633 - Data Mining for Business Applications

Winter 2020

Jasmine Li

Rabia Otry

Edna Pinzon

Anqi Zhang

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

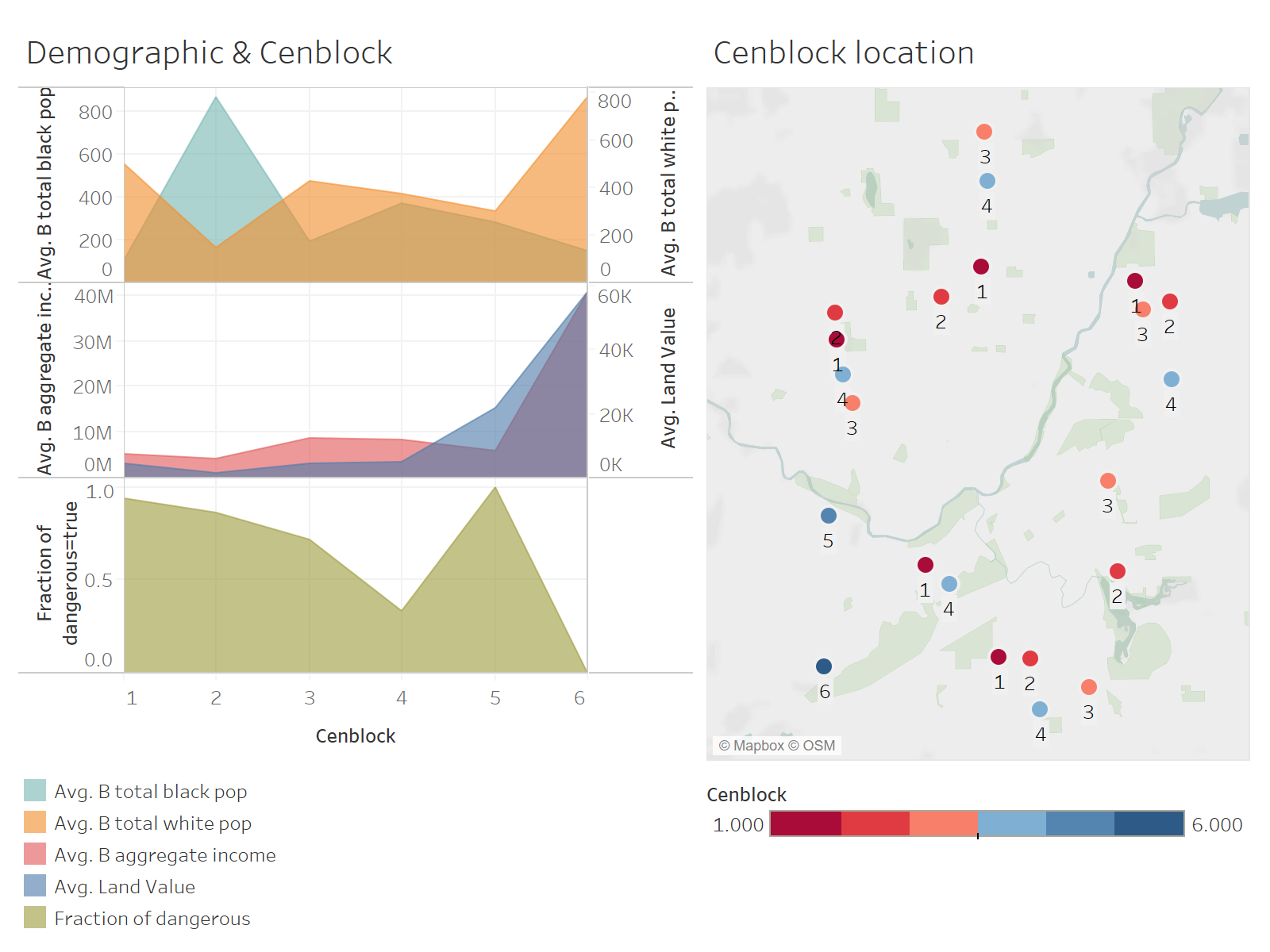
**Background:** In Flint’s water crisis problem, we are given 340 records from 2016 to create a predictive model to determine whether the water is dangerous. We used JMP and Tableau software to determine a model, create a prediction analysis, and visualize the model and demographics.

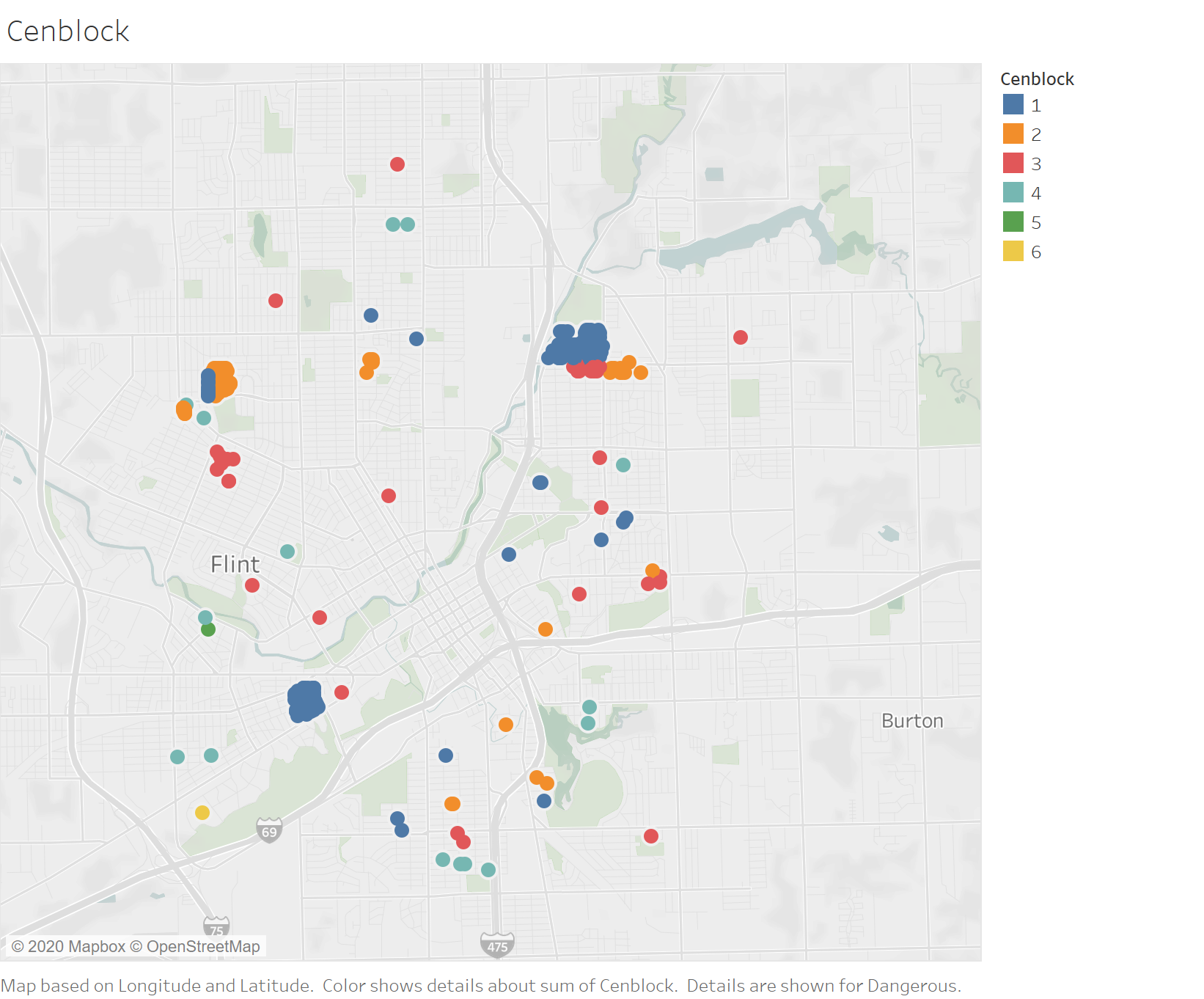
Part 1 Visualizations and their story

1. Demographics

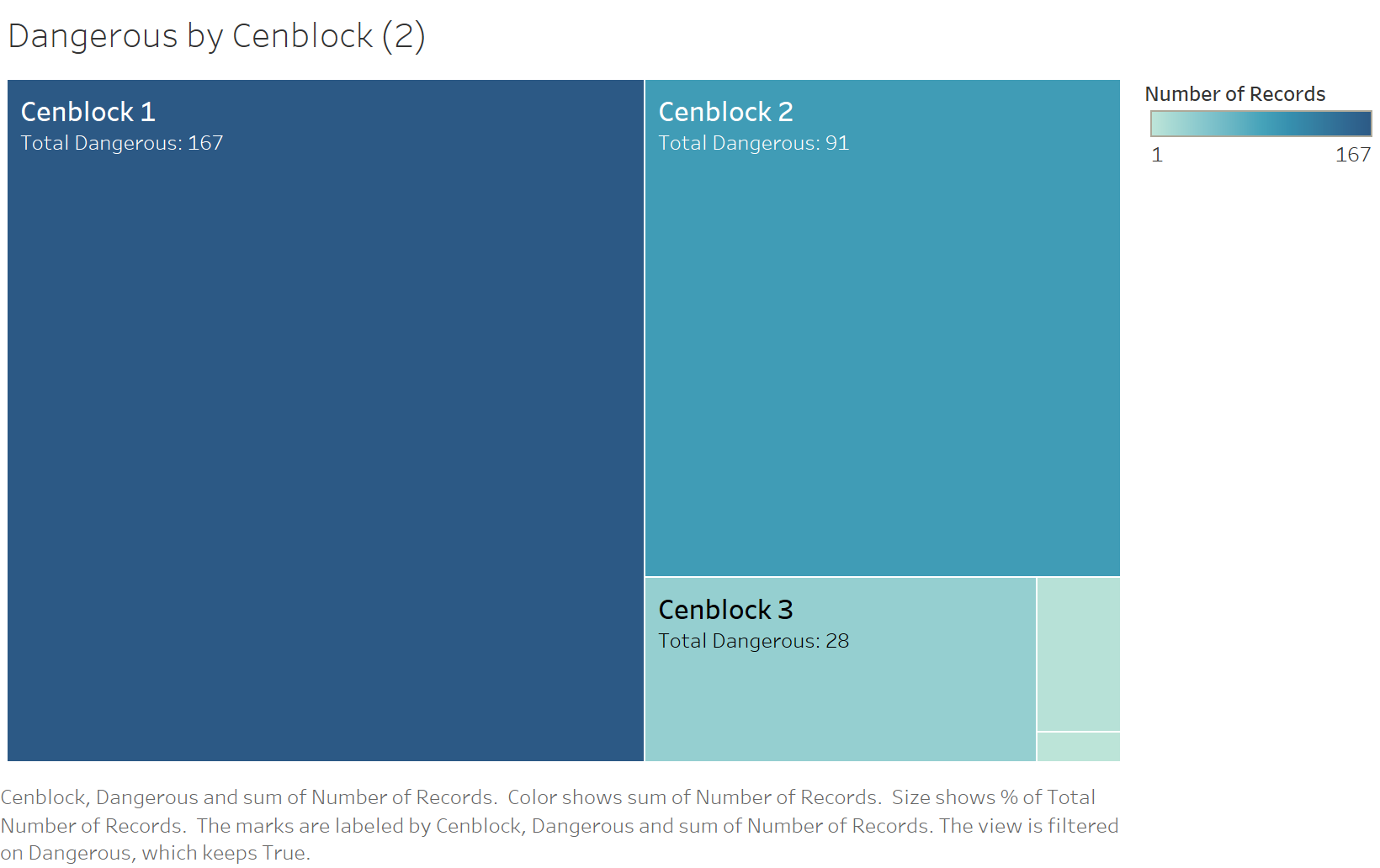
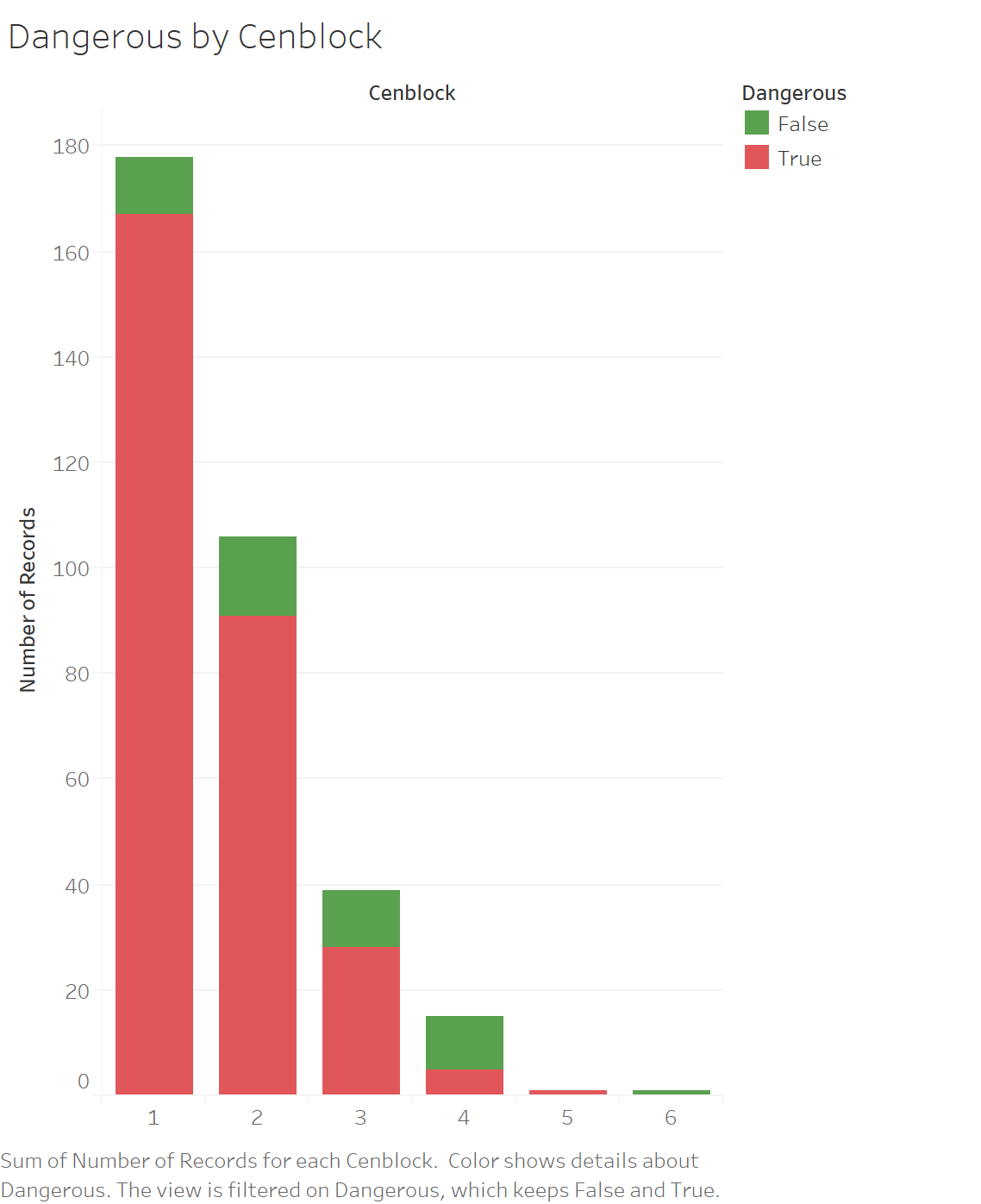
According to the American Community Survey by census block (1 to 6), the black population is the most concentrated in census block two and the white population is the most concentrated in census block 6. The average aggregate income and land value of census block 6 is the highest among all census blocks. Census block 5 has the highest fraction of dangerous service lines, with 100% service lines made of lead or being galvanized. Whereas 0% of service lines in census block 6 are dangerous.

Census block is not necessarily divided by geographic locations. Census blocks 1-4 are quite dispersed. Census block 5, however, is the closest to the Flint River, and census 6 is further from the river.



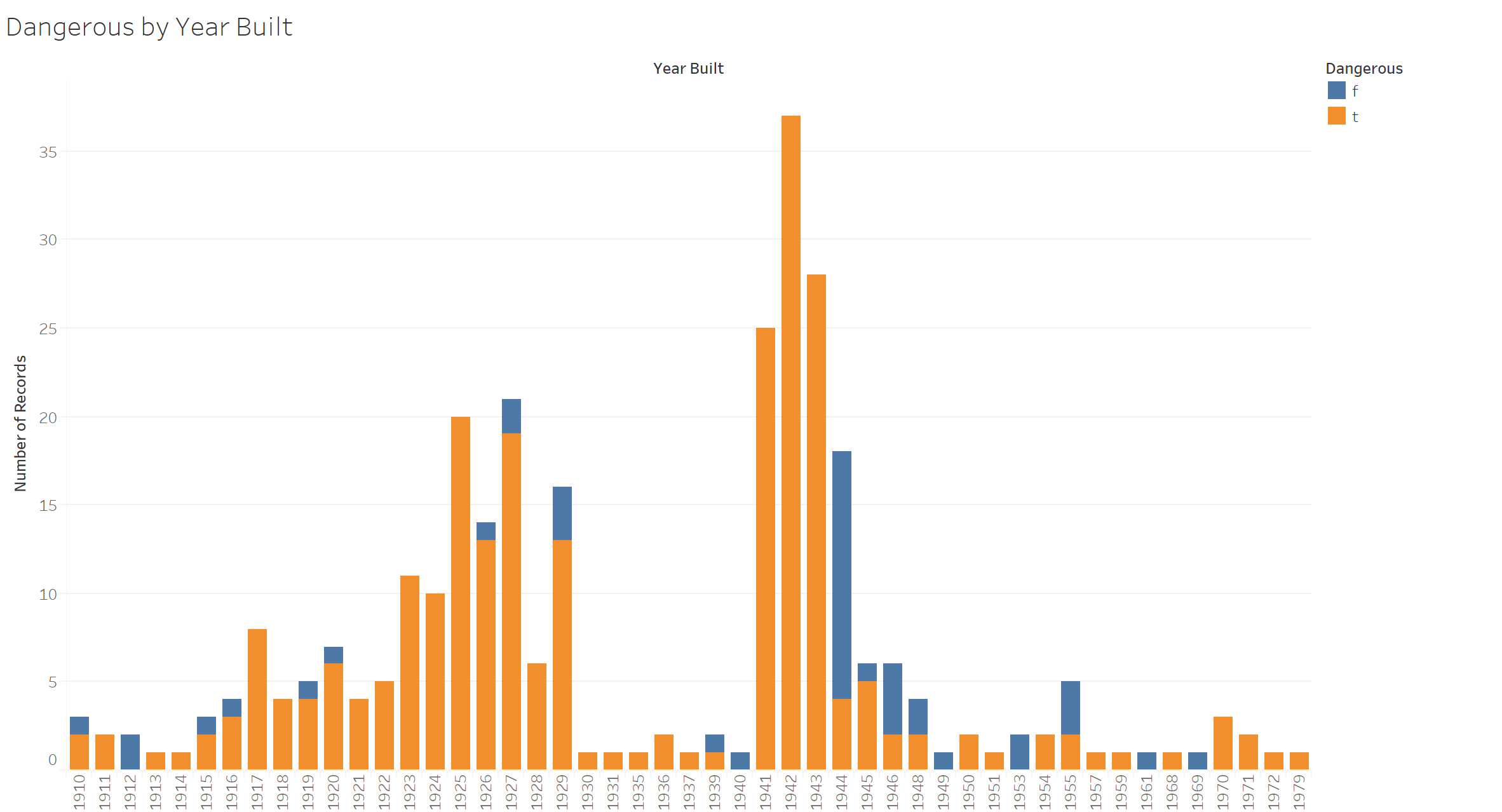


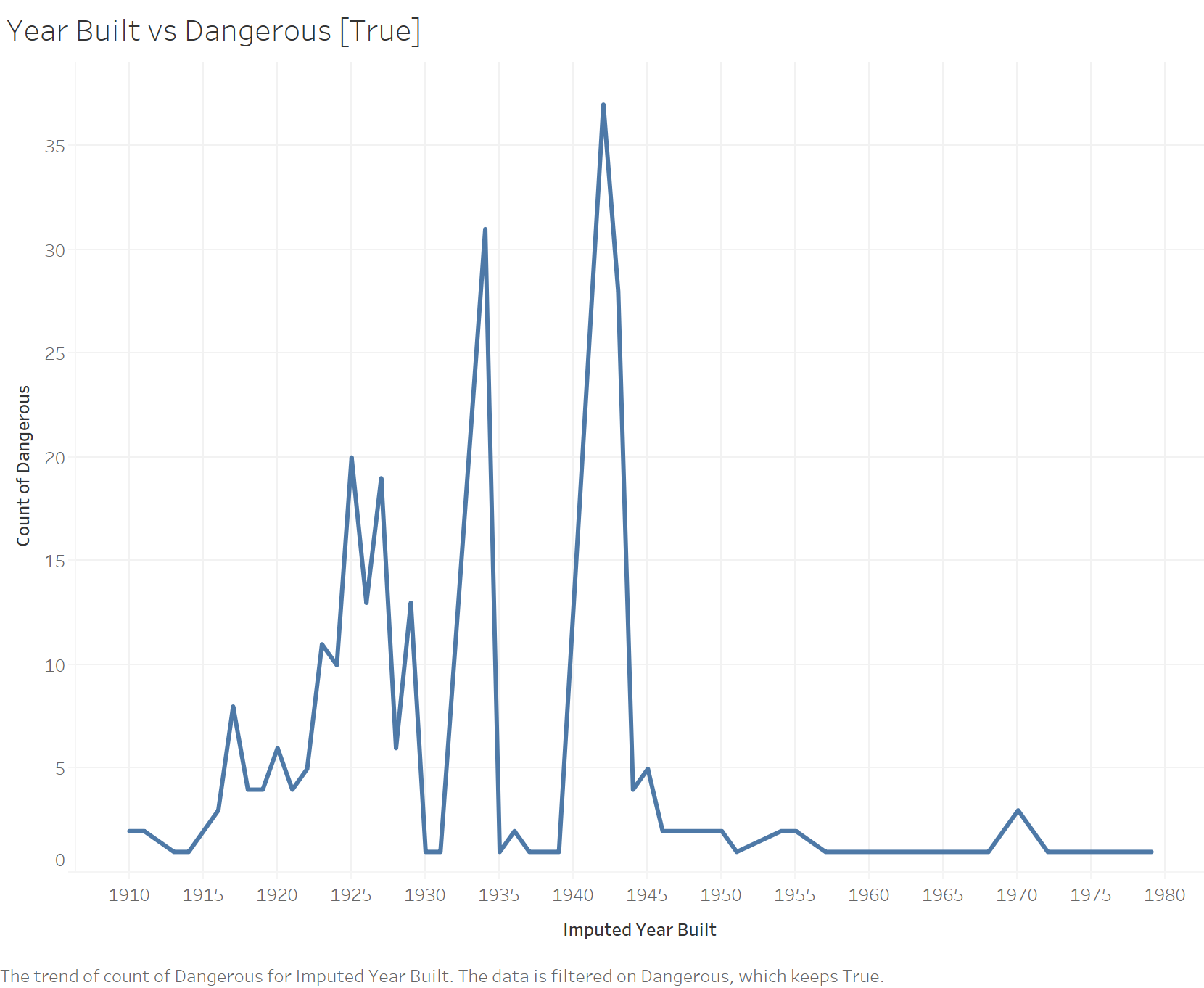
Cenblock 1 has the majority of dangerous instances, comprising nearly 60% of all dangerous cases across all cenblocks.



2. Age of property

There are two surges of building properties: after around World War One and World War Two. The service lines contained dangerous materials in most years.

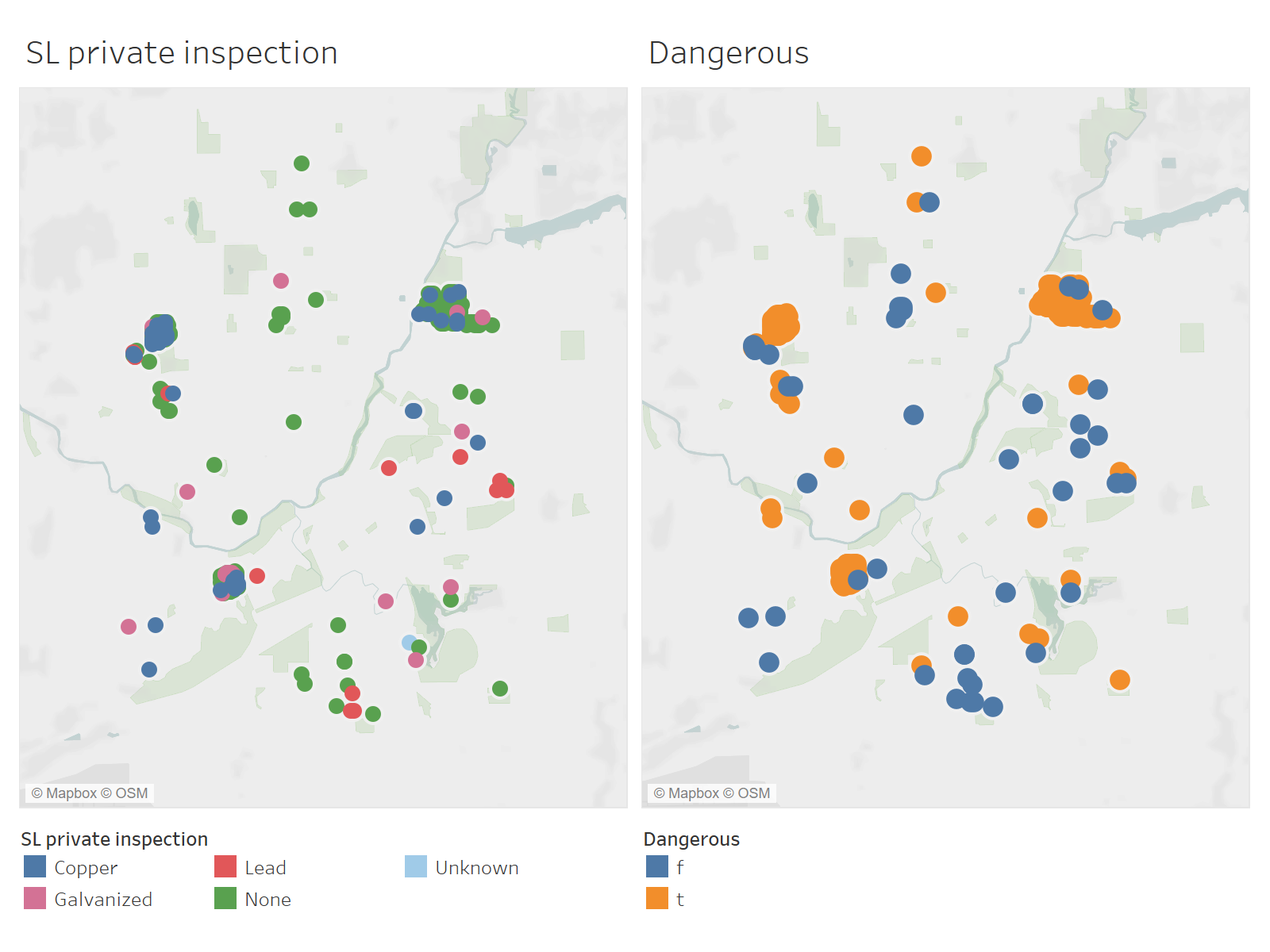




3. Service line inspection

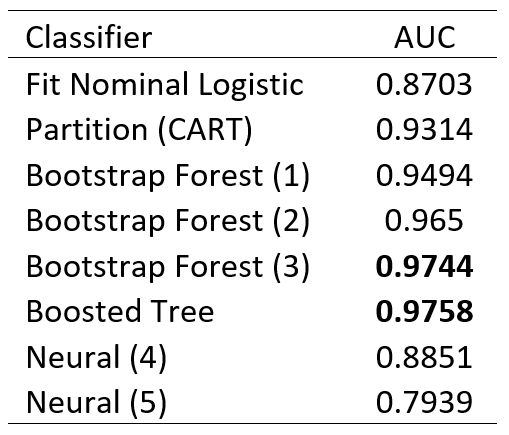
The map on the left shows the result of the inspection of private service lines conducted by Michigan Dept of Env. Quality. The result is NOT consistent with the map on the right, where the orange dots means there are lead or galvanized in either the private service line or the public service line of the property.

Some orange dots on the right map shows “unharmful” on the left map, which can be explained that there might be dangerous material in the public service line. However, the properties which are marked “Lead” or “Galvanized” on the left map and blue on the right map can be really problematic.



Part 2: Main modeling assumptions/approach

We assume that the variables given to us could be used to predict the dangerous level of service lines. We first exclude some columns that we think are not appropriate to be included in the model. We also did multivariate to check the correlation of variables. And based on that, some columns are excluded. We also assume that the best model can be selected based on AUC of the ROC. We tried several models for the prediction. Then we tried to fit different models with variables. We dynamically evaluate the variables, so then some of the variables may be included back. Here is the summary result.



(1) With continuous Imputed\_Year\_Build, defult specification

(2) With Imputed\_Year\_Build Binned, defult specification

(3) With Imputed\_Year\_Build Binned, Number of Trees in the Forest = 500 in specification

(4) With Imputed\_Year\_Build Binned, both layer 1 and layer 2 set function = TanH, node = 3

(5) With Imputed\_Year\_Build Binned, layer 1 function = TanH, node = 3; layer 2 function = Linear, node = 3

We found that Bootstrap (3) and Boosted Tree have the highest AUC. Please be noted that since we did not set up random seed for the models we discussed above, so every time, with the model running randomly, we could get slightly different AUC and ROC. However, we choose Bootstrap to be the selected prediction model. Short summary of the two methods as below. The two models mentioned above with similar results are summarized below.

Fit Bootstrap Forest

Do a Bootstrap with continuous Imputed\_Year\_Build, we got ROC = 0.9261, using default specifications. In this case, we also tried Bootstrapping with Imputed\_Year\_Build Binned, using default specifications, got an ROC = 0.9304. We also tried Bootstrapping with Imputed\_Year\_Build Binned, using Number of Trees in the Forest = 500, specifications, ROC = 0.9641.

We also tried Bootstrap with Imputed\_Year\_Build Binned, using Number of Trees in the Forest = 1000, specifications, ROC is no better than above.

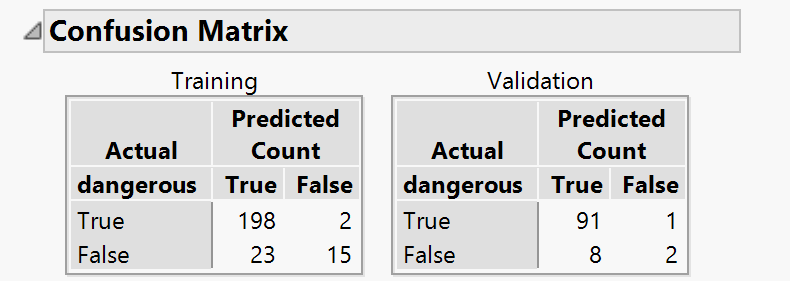
## Fit Boosted Tree Model

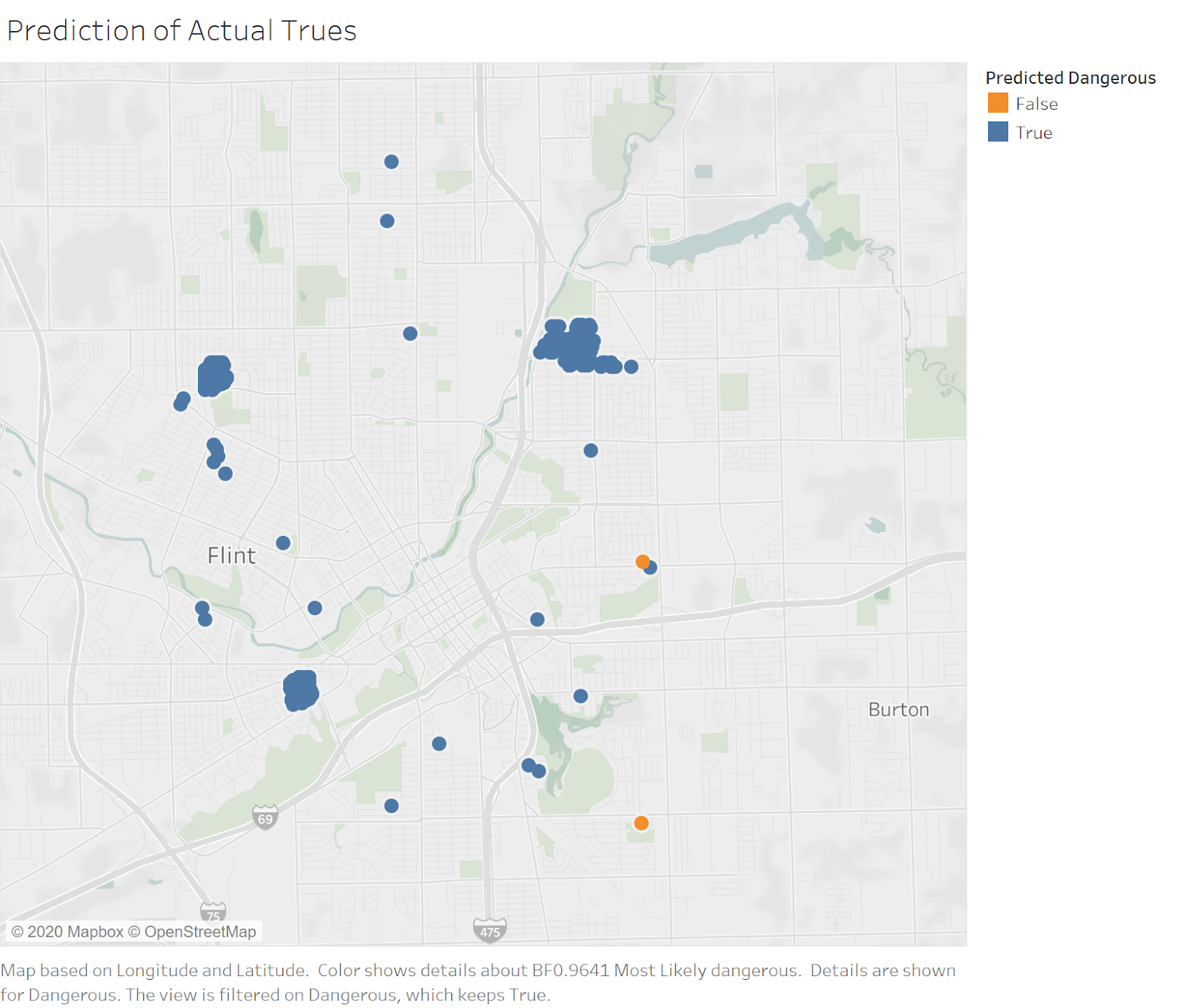
Fit Boosted Tree with continuous Imputed\_Year\_Build, number of layers = 100, ROC is close to Bootstrap method. We tried other boosted trees with different specification inputs, the results are no better than above.

Part 3 Final prediction model overview and AUC metrics

Here is a summary of all the models we tried. Using AUC of ROC, as shown below. Please be noted that since we did not set up random seed for the models we discussed above, so every time, with the model running randomly, we could get slightly different AUC and ROC. However, we still choose Bootstrap as the selected model. See model comparison and AUC in Appendix 7. Thus, we choose Bootstrapping using Imputed\_Year\_Build Binned, Number of Trees in the Forest = 500 in specification as the best choice, with ROC = 0.9641.

Visualization of our model’s predictions of actual Dangerous [True]. As you can see, our model classified almost all Trues as True.





This is also supported by our prediction model’s confusion matrix. Thus, our model predicts nearly all the accurate trues successfully, with the human health cost of misclassification of actual Trues being much higher than that of Falses.